CLAIMS

What is claimed is:

1. A method for transforming a hypermedia document containing main content and auxiliary data, the method comprising:

converting the hypermedia document into a string containing a plurality of first values and a plurality of second values, the plurality of first values corresponding to a plurality of formatting code segments within the hypermedia document and the plurality of second values corresponding to a plurality of text segments within the hypermedia document;

applying a low-pass filter to the string containing the plurality of first values and the plurality of second values; and

determining location of the main content within the hypermedia document using an output of the low-pass filter.

2. The method of claim 1 further comprising:

coding the main content in a mobile device language for display on a mobile device.

3. The method of claim 1, wherein the hypermedia document is a file written in any one of a hypertext markup language (HTML), a dynamic

HTML, an extensible HTML (XHTML), an extensible markup language (XML), JavaScript, and Visual Basic (VB) script.

4. The method of claim 1, wherein converting the hypermedia document further comprises:

parsing the hypermedia document to identify the plurality of formatting code segments and the plurality of text segments within the hypermedia document;

assigning a first value to each character within the plurality of formatting code segments; and

assigning a second value to each character within the plurality of text segments.

5. The method of claim 4 further comprising truncating a length of one of the plurality of formatting code segments when the length of said one of the plurality of formatting code segments exceeds a threshold tag length value.

6. The method of claim 1, wherein each of the plurality of first values is equal to zero.

7. The method of claim 1, wherein each of the plurality of second values is equal to one.

8. The method of claim 1, wherein the low-pass filter is a moving average filter.

9. The method of claim 8, wherein the output of the low-pass filter represents a distribution of text density over the hypermedia document.

10. The method of claim 9, wherein determining the location of the main content further comprises:

searching an output of the low-pass filter to find a position of a central peak corresponding to the highest text density within the hypermedia document; and

determining a starting position of a high text density area and an ending position of the high text density area using the position of the central peak and a threshold text density value.

11. The method of claim 10, wherein the threshold text density value is determined empirically.

12. The method of claim 1 further comprising:

varying the second value for one of the plurality of text segments based upon a weight associated with said one of the plurality of text segments.

13. The method of claim 1, wherein applying the low-pass filter further comprises:

applying a median filter to the string containing the plurality of first values and the plurality of second values to suppress high frequency signal oscillations associated with the string; and

applying a moving average filter to an output of the median filter to combine a plurality of closely spaced text segments contained in the output of the median filter into a set of larger text segments.

14. The method of claim 13, wherein determining the location of the main content further comprises:

applying a rising and falling edge detector to an output of the median filter to identify the largest reasonably contiguous text segment within the set of larger segments.

15. The method of claim 14, wherein the largest reasonably contiguous text segment is identified using a threshold text value.

16. An apparatus for transforming a hypermedia document containing main content and auxiliary data, the apparatus comprising:

a converter to convert the hypermedia document into a string containing a plurality of first values and a plurality of second values, the plurality of first values corresponding to a plurality of formatting code segments within the hypermedia document and the plurality of second values corresponding to a plurality of text segments within the hypermedia document;

a low-pass filter to apply to the string containing the plurality of first values and the plurality of second values; and

a location calculator to determine location of the main content within the hypermedia document using an output of the low-pass filter.

17. The apparatus of claim 16 further comprising:

an encoder to code the main content in a mobile device language for display on a mobile device.

18. The apparatus of claim 16, wherein the hypermedia document is a file written in any one of a hypertext markup language (HTML), a dynamic HTML, an extensible HTML (XHTML), an extensible markup language (XML), JavaScript, and Visual Basic (VB) script.

19. The apparatus of claim 16 further comprising a parser to identify the plurality of formatting code segments and the plurality of text segments within the hypermedia document.

20. The apparatus of claim 16 wherein the converter is to convert the hypermedia document by assigning a first value to each character within the plurality of formatting code segments and assigning a second value to each character within the plurality of text segments.

21. The apparatus of claim 20 wherein the converter is to truncate a length of one of the plurality of formatting code segments when the length of said one of the plurality of formatting code segments exceeds a threshold tag length value.

22. The apparatus of claim 16, wherein each of the plurality of first values is equal to zero.

23. The apparatus of claim 16, wherein each of the plurality of second values is equal to one.

24. The apparatus of claim 16, wherein the low-pass filter is a moving average filter.

25. The apparatus of claim 24, wherein the output of the low-pass filter represents a distribution of text density over the hypermedia document.

26. The apparatus of claim 25, wherein the location calculator is to determine the location of the main content by searching an output of the low-pass filter to find a position of a central peak corresponding to the highest text density within the hypermedia document, and by determining a starting position of a high text density area and an ending position of the high text density area using the position of the central peak and a threshold text density value.

27. The apparatus of claim 1 wherein the converter is to vary the second value for one of the plurality of text segments based upon a weight associated with said one of the plurality of text segments.

28. The apparatus of claim 16, wherein the low-pass filter further comprises:

a median filter to be applied to the string containing the plurality of first values and the plurality of second values to suppress high frequency signal oscillations associated with the string; and

a moving average filter to be applied to an output of the median filter to combine a plurality of closely spaced text segments contained in the output of the median filter into a set of larger text segments.

29. The apparatus of claim 28, wherein the location calculator is to determine the location of the main content by applying a rising and falling edge detector to an output of the median filter to identify the largest reasonably contiguous text segment within the set of larger segments.

30. The apparatus of claim 29, wherein the location calculator is to identify the largest reasonably contiguous text segment using a threshold text value.

31. A medium readable by a machine, the medium having stored thereon a sequence of instructions which, when executed by the machine, cause the machine to:

convert the hypermedia document into a string containing a plurality of first values and a plurality of second values, the plurality of first values corresponding to a plurality of formatting code segments within the hypermedia document and the plurality of second values corresponding to a plurality of text segments within the hypermedia document;

apply a low-pass filter to the string containing the plurality of first values and the plurality of second values; and

determine location of the main content within the hypermedia document using a low-pass filter output.

32. A method for transforming a web page containing main content and auxiliary data, the method comprising:

converting the web page into a string containing a plurality of first values and a plurality of second values, the plurality of first values corresponding to a plurality of formatting code segments within the web page and the plurality of second values corresponding to a plurality of text segments within the web page;

applying a moving average filter to the string containing the plurality of first values and the plurality of second values to generate an output representing a distribution of text density over the web page;

searching the output of the moving average filter to find a position of a central peak corresponding to the highest text density within the web page;

determining a starting position of a high text density area and an ending position of the high text density area using the position of the central peak and a threshold text density value to determine location of the main content within the web page; and

coding the main content in a mobile device language for display on a mobile device.

33. The method of claim 32 further comprising truncating a length of one of the plurality of formatting code segments when the length of said one of the plurality of formatting code segments exceeds a threshold tag length value.

34. The method of claim 32, wherein each of the plurality of first values is equal to zero and each of the plurality of second values is equal to one.

35. The method of claim 32 further comprising:

varying the second value for one of the plurality of text segments based upon a weight associated with said one of the plurality of text segments.

36. A method for transforming a web page containing main content and auxiliary data, the method comprising:

converting the web page into a string containing a plurality of first values and a plurality of second values, the plurality of first values corresponding to a plurality of formatting code segments within the web

page and the plurality of second values corresponding to a plurality of text segments within the web page;

applying a median filter to the string containing the plurality of first values and the plurality of second values to suppress high frequency signal oscillations associated with the string;

applying a moving average filter to an output of the median filter to combine a plurality of closely spaced text segments contained in the output of the median filter into a set of larger text segments;

applying a rising and falling edge detector to an output of the median filter to identify the largest reasonably contiguous text segment within the set of larger segments using a threshold text value, the largest reasonably contiguous text segment corresponding to the main content of the web page; and

coding the main content in a mobile device language for display on a mobile device.

37. The method of claim 36 further comprising truncating a length of one of the plurality of formatting code segments when the length of said one of the plurality of formatting code segments exceeds a threshold tag length value.

38. The method of claim 36, wherein each of the plurality of first values is equal to zero and each of the plurality of second values is equal to one.

39. The method of claim 36 further comprising:

varying the second value for one of the plurality of text segments based upon a weight associated with said one of the plurality of text segments.